Vol.2020-MBL-94 No.47
Vol.2020-UBI-65 No.47
2020/3/3

ビデオからの3次元姿勢を用いた行動認識における精度向上の試み

安達 康平[1,a)]　井上 創造[1,b)]

概要：モーションキャプチャシステムを用いた行動認識は高い精度で認識ができる一方で，システムを使用する上で，準備や後処理に非常に多くの時間を要する．そこで我々は，ビデオから 3 次元姿勢を推定する手法を代わりに使用する．しかしながら，カメラの撮影位置が変化することにより認識精度が大きく下がるという問題があった．本稿ではビデオから推定した 2 次元姿勢や 3 次元姿勢で行動認識を行い，認識精度の比較を行った．さらに推定した 3 次元姿勢を回転させることで学習データのデータ増強を行い，異なる撮影位置における認識精度の改善を試みた．結果として，異なる撮影位置における認識精度が 2 次元姿勢で行動認識を行ったときより最大で 54.8%改善することがわかった．

1. はじめに

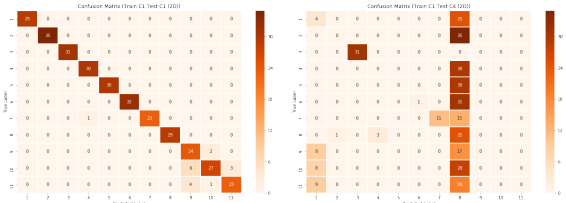
行動認識は，センサデータやビデオから人間の様々な種類の行動を認識する技術であり，ユビキタスコンピューティングの分野で盛んに研究が行われている [1].

一般的に行動認識手法は，教師あり機械学習を用いるためラベル付きデータ収集を必要とする．我々の研究では，3 次元姿勢を計測できるモーションキャプチャシステムを用いて行動認識 [8] を行っているが，モーションキャプチャを使用する上で，準備や後処理などに非常に時間がかかる点が問題である．

近年，深層学習を用いた，ビデオや 2 次元姿勢から 3 次元姿勢を推定する手法 [4], [7] が提案されており，我々は，これらをモーションキャプチャの代わりに使えないかと考えた．

そこで，我々の研究では深層学習を用いてビデオから人の 2 次元姿勢を抽出するライブラリ OpenPose [2] と 2 次元姿勢から 3 次元姿勢推定を行う深層学習フレームワーク [4] を用いて，ビデオから推定された 2 次元姿勢および 3 次元姿勢で行動認識を行った [10].

しかしながら，カメラの撮影位置が変化することにより認識精度が下がることが分かった (図 1)．本稿では，ビデオから推定された 2 次元姿勢および 3 次元姿勢で行動認識を行い，認識精度の比較を行った．さらに，推定された 3 次元姿勢をアフィン変換を用いて回転させることで学習データのデータ増強を行い，異なる撮影位置における認識精度の改善を試みた (図 2)．結果として，異なる撮影位置における認識精度が最大で 54.8%，2 次元姿勢で行動認識を行ったときより改善することがわかった



(a) F1-Score:94.8%　(b) F1-Score:18.2%

図 1 ビデオから推定した 2 次元姿勢で行動認識を行ったときの混同行列．テストデータが学習データと同じ撮影位置だった場合，高い精度で認識できる (右) が，撮影位置が異なった場合，認識精度が低くなる (左)

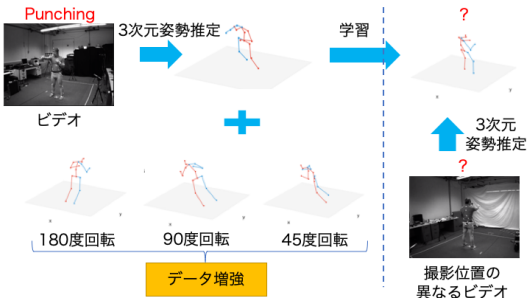


図 2 ビデオから 3 次元姿勢を推定した後．アフィン変換を用いて 3 次元姿勢を回転することで，データ増強を行い，撮影位置の異なるビデオから推定した 3 次元姿勢に対して認識しやすくする

[1] 九州工業大学
Kyushu Institute of Technology
[a)] adachi@sozolab.jp
[b)] sozo@sozolab.jp

Vol.2020-MBL-94 No.47
Vol.2020-UBI-65 No.47
2020/3/3

2. 関連研究

高嶋らは，ビデオからリアルタイムに行動解析することを目標としており，ビデオから画像 1 枚または 10 枚取得し，OpenPose で抽出された姿勢情報を説明変数として入力し，様々な機械学習手法を用いて行動認識した際の認識精度の比較を行っている [11].

Varol らは，ビデオの行動認識において，カメラに対する人の向きが変化することで認識精度が低下する問題に対して，人に対して 0 度の位置から撮影したビデオから，人が異なる角度を向いて行動するビデオを人工的に作成し，学習データを増強することで認識精度を向上させる手法を提案している [9].

また，我々の研究では，加速度センサ・モーションキャプチャ・OpenPose で抽出された 2 次元姿勢の間で転移可能な行動認識モデルの構築を行っている [6].

これらに対し，本研究ではビデオから 3 次元姿勢推定を行い，推定した 3 次元姿勢に対してアフィン変換を用いて回転させることでデータ増強し，異なる撮影位置における認識精度の向上を試みる．

3. 実験方法

本稿では，ビデオから推定した 2 次元姿勢と 3 次元姿勢，データ増強された 3 次元姿勢から特徴量抽出を行い，それぞれで行動認識を行った結果を比較する．実験の概要を図 3 に示す．

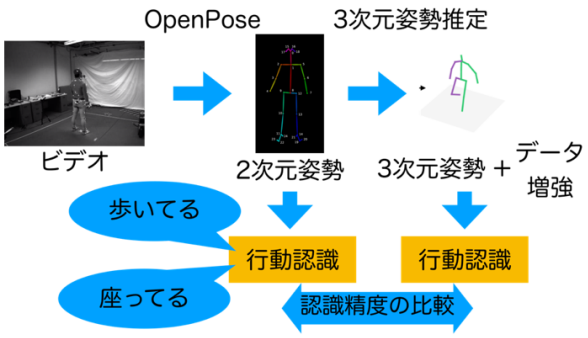


図 3 ビデオから OpnePose を用いて 2 次元姿勢を推定し，2 次元姿勢からさらに 3 次元姿勢を推定をする．その後，2 次元および 3 次元姿勢，データ増強を行った 3 次元姿勢を用いて行動認識を実施して認識精度を比較する．

3.1 データセット

実験に使用するデータセットには，Berkeley MHAD(Multimodal Human Action Database)[5] を用いた．このデータセットには，7 名の男性と，5 名の女性の計 12 名の被験者が，表 1 に示す 11 つの行動を各行動につき 5 回ずつ音声・ビデオ・加速度センサ・モーションキャプチャで取ったものが含まれている．我々は，本実験のためにカメラから撮影されたビデオを用いた．このビデオは被験者を囲うように設置されたカメラから撮影されたものであり，図 4 にカメラの配置図，図 5 に各カメラから撮影された例を示す．

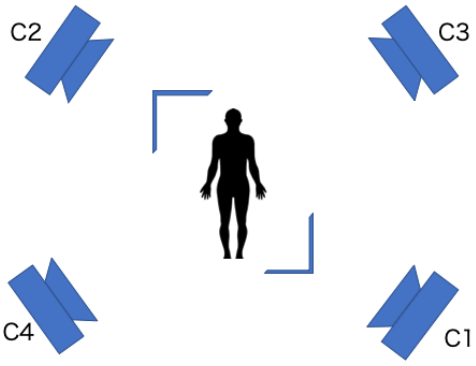


図 4 動画を撮影するカメラの配置図．被験者を囲うように 4 台設置されている．被験者は C1,C4 のカメラの方面を向いて行動を行う．

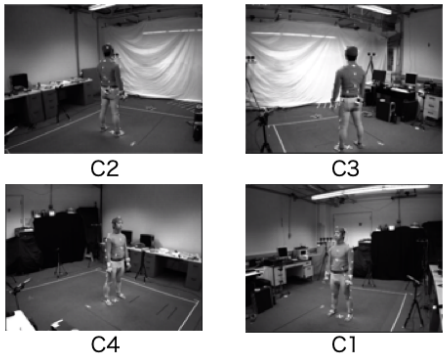


図 5 各カメラからの撮影する動画の一例．C1,C4 は被験者を前面を写し，C2,C3 は被験者の後ろ姿を映す．

表 1 被験者が行動する行動の内容．全部で 11 種類ある．

クラス	行動
1	Jumping in place
2	Jumping jacks
3	Bending - hands up all the way down
4	Punching(boxing)
5	Waving - two hands
6	Waving - one hand(right)
7	Clapping hands
8	Throwing a ball
9	Sit down then stand up
10	Sit down
11	Stand up

Vol.2020-MBL-94 No.47
Vol.2020-UBI-65 No.47
2020/3/3

3.2 OpenPose を用いたビデオから 2 次元姿勢の抽出

OpenPose を用いて各動画の各フレームに対して 2 次元の 25 個のキーポイントを取得する．抽出されたキーポイントの例を図 6 に示す．

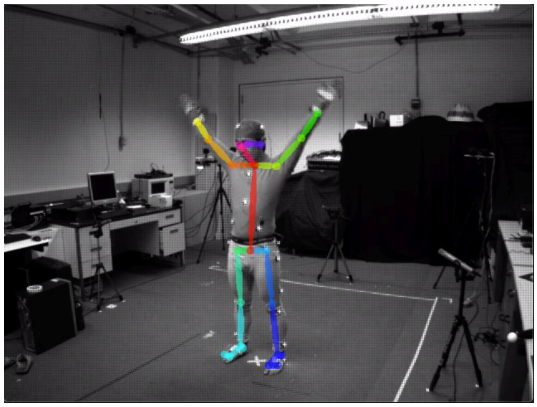


図 6 OpenPose によって取得された 25 個のキーポイント

3.3 抽出された 2 次元姿勢から 3 次元姿勢を推定

3.2 で抽出されたキーポイントから，3 次元姿勢を推定する．推定には， [4] の深層学習を用いたフレームワークを用いる．これは，Dropout, Batch Normalization と ReLU, Redisual Connection を用いて，予測した姿勢座標と実際の姿勢座標の差が最小になるよう学習された深層学習モデルである．実装には，事前学習済みモデルと [4] を OpenPose の出力[*1]を入力として使えるように改良した 3d-pose-baseline[*2]を用いて，3 次元の 32 個のキーポイントを推定する．3 次元姿勢推定を行った例を図 7 に示す．

図 7 3d-pose-baseline で推定された 3 次元姿勢

[*1] https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/doc/output.md
[*2] https://github.com/ArashHosseini/3d-pose-baseline

3.4 3 次元アフィン変換を用いたデータ増強

推定した 3 次元姿勢に対してアフィン変換を用いて鉛直方向である Z 軸回りに回転させて，データ増強を行う．回転に用いる式は次式で表され，実際に 3 次元姿勢を回転させた例を図 8 に示す．データ増強は，推定した 3 次元姿勢に対して，+45°,+90°,+180° をそれぞれ回転されたものと，それらを 2 つずつ組み合わせて回転させたものを学習データに追加するように増強した．従ってデータ増強の方法は 6 種類となり，表 2 にデータ増強の方法を示す．

$$R_z = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1}$$

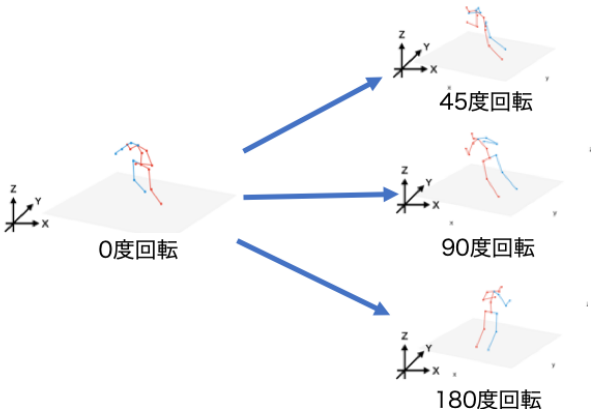


図 8 推定した 3 次元姿勢に対して，アフィン変換を用いて Z 軸回りに回転させる

表 2 データ増強のための回転パターン

	回転パターン
1	45°
2	90°
3	180°
4	45°, 90°
5	45°, 180°
6	90°, 180°

3.5 特徴量抽出

3.2 の 2 次元姿勢および 3.3 の 3 次元姿勢から特徴量を抽出する．特徴量抽出には，各動画から抽出された時系列に並ぶ各キーポイント座標値の最大値・最小値・標準偏差・平均を使用する．

従って，各キーポイントから抽出される特徴量数は，2 次元姿勢では 8 個，3 次元姿勢では 12 個となり，各次元姿勢における特徴量数は 2 次元姿勢で 200 個，3 次元姿勢で 384 個となる．

Vol.2020-MBL-94 No.47
Vol.2020-UBI-65 No.47
2020/3/3

3.6 機械学習を用いた分類モデルの作成

3.5 にて得られた特徴量を説明変数として入力し，行動を識別する分類モデルを作成する．分類モデルの作成には機械学習アルゴリズムの RandomForest [3] を用いた．

4. 評価手法

データセットの行動データの内，5 割を学習データ，残りの 5 割をテストデータに分け，学習データおよびテストデータを撮影位置の異なるものにして評価した．このとき，異なる撮影位置であり，且つ同じタイミングで記録された行動が学習データとテストデータの両方に含まれないように注意した．また，比較として学習データ，テストデータを同撮影位置にした場合の評価も行った．評価指標として，テストデータに対する F1-Score を用いた．

これらを 2 次元姿勢，3 次元姿勢，データ増強した 3 次元姿勢に対して行った．

5. 結果

表 3 に 2 次元姿勢・3 次元姿勢・データ増強を行った 3 次元姿勢に対して行動認識を行った結果を示す．学習データおよびテストデータで**同じ撮影位置のカメラを使用した場合**と**異なる撮影位置のカメラを使用した場合**の結果の比較を行う．

5.1 同じ撮影位置のカメラを使用していた場合

表 3 の学習データが C1，テストデータが C1 などのように同じ撮影位置のカメラを使用した場合の結果から 2 次元姿勢，3 次元姿勢共に高い精度で認識できることが分かった．また，学習データに回転させた 3 次元姿勢をデータ増強として加えたときでも高い認識で認識できることが分かった．

5.2 異なる撮影位置のカメラを使用していた場合 (データ増強なし)

表 3 の学習データが C1，テストデータが C2 などのように異なる撮影位置のカメラを使用した場合の結果から，2 次元姿勢が 3 次元姿勢よりも認識精度が高かったのは，12 ケース中 11 ケースであり，従ってこの場合 3 次元姿勢よりも 2 次元姿勢の方が認識精度が高いことが分かった．

3 次元姿勢が 2 次元姿勢よりも認識精度が高かったのは，学習データが C1，テストデータが C4 のカメラをを用いたケースで，15.3%認識精度が改善し，3 次元姿勢が 2 次元姿勢よりも認識精度が低かったのは，学習データが C3，テストデータが C1 のカメラを用いたケースで 38.6%認識精度が悪化した．

5.3 異なる撮影位置のカメラを使用していた場合 (データ増強あり)

表 3 の学習データが C1，テストデータが C2 で学習データに対してデータ増強を加えたケースなどのように，異なる撮影位置のカメラを使用，且つ学習データに対してデータ増強を行った結果から，異なる撮影位置において 3 次元姿勢で学習データに対してデータ増強をした際に，2 次元姿勢の認識精度より改善したものは 12 ケース中 10 ケースであった．図 9,10 には，最も認識精度が改善した例を示す．これは，学習データが C1，テストデータが C4 のカメラを用いたときであり，C1 の学習データの 3 次元姿勢を 45°,180° の回転させてデータ増強として加えたときに，2 次元姿勢よりも 54.8%認識精度が改善した．

反対に，図 11,12 に，認識精度が悪化した例も示す．これは学習データが C3，テストデータが C1 のカメラを用いたときであり，C3 の学習データの 3 次元姿勢を 45° 回転させてデータ増強として加えたときに 22.8%認識精度が悪化した．

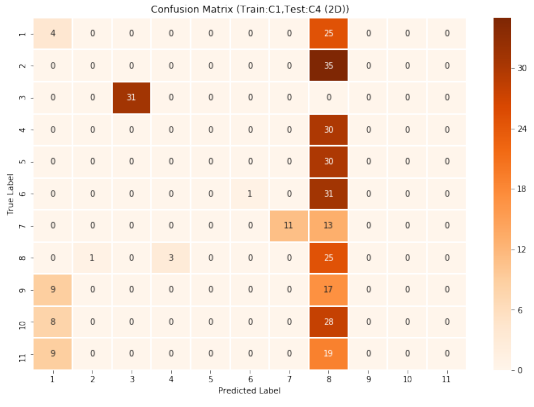


図 9 学習データ:C1, テストデータ:C4 のカメラから推定した 2 次元姿勢を用いて行動認識を行った際の混同行列 (F1-Score:18.2%)

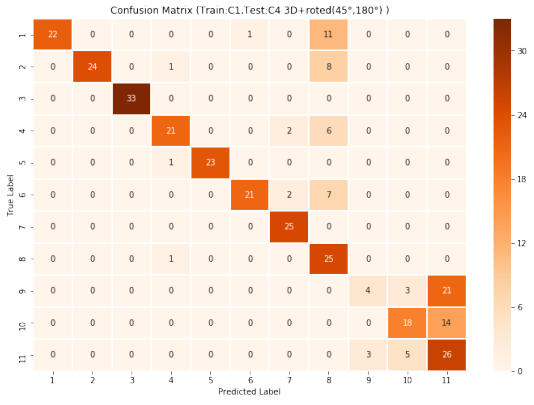


図 10 学習データ:C1, テストデータ:C4 のカメラから推定した 3 次元姿勢と 45°，180° 回転させた姿勢でデータ増強した際の混同行列 (F1-Score:73.9%)

表 3 2 次元姿勢・3 次元姿勢・データ増強した 3 次元姿勢姿勢に対して行動認識を実施したときのテストデータに対する F1-Score の結果 (%). 縦軸の学習データが学習に用いたカメラ，横軸のテストデータがテストに用いたカメラを表す．C1，C2，C3，C4 はカメラの撮影位置を表し，C1，C4 のカメラは人の前面を映し，C2，C3 のカメラは人の後面を映す．

		テストデータ			
		C1	C2	C3	C4
学習データ	C1(2D)	94.8	**39.1**	46.4	18.2
	C1(3D)	**96.4**	20.0	28.1	33.5
	C1(+45°)	95.3	32.7	41.1	62.7
	C1(+90°)	96.4	27.5	40.3	62.8
	C1(+180°)	93.9	37.4	**47.5**	68.7
	C1(+45°, +90°)	93.6	30.3	40.4	41.0
	C1(+45°, +180°)	94.8	35.7	42.6	**73.9**
	C1(+90°, +180°)	95.6	31.4	44.2	72.5
学習データ	C2(2D)	**58.0**	**93.5**	57.3	41.8
	C2(3D)	26.1	93.2	34.0	20.5
	C2(+45°)	43.7	91.5	65.5	39.1
	C2(+90°)	45.6	90.7	64.7	36.2
	C2(+180°)	56.2	91.2	**77.8**	**51.6**
	C2(+45°, +90°)	46.4	90.0	69.1	41.8
	C2(+45°, +180°)	45.7	89.8	70.9	44.9
	C2(+90°, +180°)	46.4	90.7	69.1	41.8
学習データ	C3(2D)	40.0	47.4	94.3	27.1
	C3(3D)	1.37	28.2	**94.8**	1.29
	C3(+45°)	17.2	42.5	94.5	7.4
	C3(+90°)	34.5	45.8	93.3	32.4
	C3(+180°)	**40.5**	44.5	93.0	**40.8**
	C3(+45°, +90°)	36.6	47.0	93.3	34.0
	C3(+45°, +180°)	36.1	**49.6**	93.3	37.2
	C3(+90°, +180°)	35.4	47.4	93.9	39.8
学習データ	C4(2D)	54.4	26.7	11.4	93.0
	C4(3D)	53.3	7.34	4.82	**96.6**
	C4(+45°)	71.4	15.1	11.5	**96.6**
	C4(+90°)	78.1	45.4	46.1	95.9
	C4(+180°)	72.7	45.7	47.6	94.4
	C4(+45°, +90°)	**86.8**	47.2	44.2	95.0
	C4(+45°, +180°)	77.1	42.3	46.1	95.9
	C4(+90°, +180°)	84.1	**47.8**	**49.1**	95.2

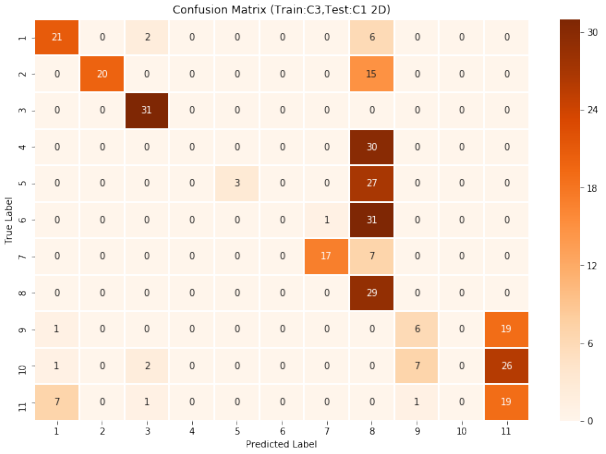


図 11 学習データ:C3, テストデータ:C1 のカメラから推定した 2 次元姿勢を用いて行動認識を行った際の混同行列 (F1-Score:40.0%)

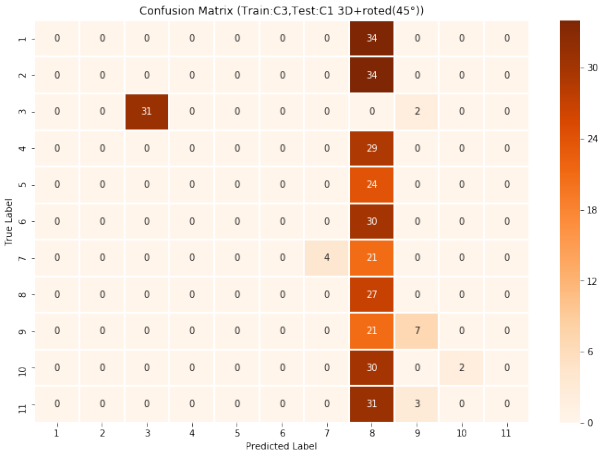
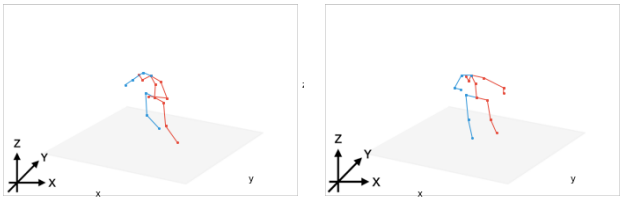
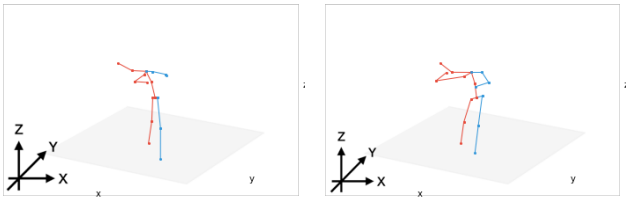


図 12 学習データ:C3, テストデータ:C1 のカメラから推定した 3 次元姿勢と 45° 回転させた姿勢でデータ増強した際の混同行列 (F1-Score:17.2%)



(a) C1 のカメラ　(b) C4 のカメラ

図 13 C1，C4 のカメラから撮影されたビデオから推定した 3 次元姿勢



(a) C2 のカメラ　(b) C3 のカメラ

図 14 C2，C3 のカメラから撮影されたビデオから推定した 3 次元姿勢

6. 考察

6.1 データ増強を行ったことによる認識精度の改善

学習データに C1 のカメラを用いるとき推定した 3 次元姿勢に対して 45°，180° の回転をさせたものを学習データに増強し，テストデータを C4 のカメラを用いたときに 2 次元姿勢よりも認識精度が 54.8%改善した．従って，3 次元姿勢を回転させることによって異なる撮影位置における 3 次元姿勢でも正しく分類可能である分類器のモデル作成に成功したと言える．

特に，C1 と C4 のカメラの組み合わせにおけるデータ増強をしたときの認識精度の改善が見られた．理由として，C1 と C4 はカメラの人の前面を撮影しており (図 4)，後面からの撮影よりも 3 次元姿勢を行ったときの誤差が少なく，3 次元姿勢を回転するデータ増強が誤差や誤検知などの影響を受けずに上手くいったと考えられる (図 13)．

6.2 データ増強を行ったことによる認識精度の悪化

学習データに C3 のカメラを用いるとき推定した 3 次元姿勢に対して 45° 回転させたもの学習データに増強し，テストデータを C4 のカメラのを用いたとき，2 次元姿勢よりも認識精度が 22.8%と大きく下がることも分かった．

また，人の後面から撮影するカメラ (C3，C2) に関しては，学習データにおけるデータ増強をした際の認識精度の改善が人の前面を撮影するカメラ (C1，C4) と比べて低いことが分かる．理由として，後面から人を映すカメラから推定した 3 次元姿勢の誤差や検知が，前面から人を映すカメラに比べて大きくそれらが影響していると考えられる (図 14)．さらに，表 3 の結果から，学習データとテストデータが前面同士や後面同士のデータ増強したときにおける認識精度の改善は見られるが，そうでない場合，認識精度の改善が低い．これらの理由も前述した，後面から人を映すカメラから推定した 3 次元姿勢の誤差や検知が影響していると考えられる．

従って，3 次元姿勢を用いたデータ増強する際には最適な回転角や推定した 3 次元姿勢の誤差などを考慮する必要あると言える．

7. まとめと今後の課題

本稿では，ビデオから推定した 3 次元姿勢で行動認識を行う際に，3 次元姿勢に対してアフィン変換を用いて回転させた姿勢を学習データに追加することでデータ増強を行い，学習データがテストデータと異なる撮影位置の場合における認識精度の改善を試みた．結果として，異なる撮影位置における認識精度が最大で 54.8%向上することが分かった．また，データ増強の回転角によっては精度が下がることも分かった．今後の課題として，3 次元姿勢を用いたデータ増強の際の最適な回転角，分類モデルに悪影響を与えないための増強データの選択手法，異なる撮影位置対してさらにロバスト性を高めるための特徴量抽出方法の提案などが挙げられる．

参考文献

[1] Bulling, A., Blanke, U. and Schiele, B.: A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors, *ACM Comput. Surv.*, Vol. 46, No. 3 (online), DOI: 10.1145/2499621 (2014).

[2] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. and Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields, *arXiv preprint arXiv:1812.08008* (2018).

[3] Liaw, A. and Wiener, M.: Classification and Regression by RandomForest, *Forest*, Vol. 23 (2001).

[4] Martinez, J., Hossain, R., Romero, J. and Little, J. J.: A simple yet effective baseline for 3d human pose estimation, *ICCV* (2017).

[5] Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R. and Bajcsy, R.: Berkeley MHAD: A comprehensive Multimodal Human Action Database., *WACV*, IEEE Computer Society, pp. 53–60 (online), available from ⟨http://dblp.uni-trier.de/db/conf/wacv/wacv2013.htmlOfliCKVB13⟩ (2013).

[6] Okita, T. and Inoue, S.: Activity Recognition: Translation across Sensor Modalities Using Deep Learning, *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, New York, NY, USA, Association for Computing Machinery, pp. 1462–1471 (online), DOI: 10.1145/3267305.3267512 (2018).

[7] Pavllo, D., Feichtenhofer, C., Grangier, D. and Auli, M.: 3D human pose estimation in video with temporal convolutions and semi-supervised training, *Conference*

on Computer Vision and Pattern Recognition (CVPR) (2019).

[8] Takeda, S., Lago, P., Okita, T. and Inoue, S.: Reduction of Marker-Body Matching Work in Activity Recognition Using Motion Capture, *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, New York, NY, USA, Association for Computing Machinery, pp. 835–842 (online), DOI: 10.1145/3341162.3345591 (2019).

[9] Varol, G., Laptev, I., Schmid, C. and Zisserman, A.: Synthetic Humans for Action Recognition from Unseen Viewpoints (2019).

[10] 安達康平，大北　剛，井上創造：ビデオからの 3 次元姿勢推定と機械学習を用いた行動認識の試み，日本知能情報ファジィ学会九州支部学術講演会予稿集，pp. 40–43 (2019).

[11] 高崎智香子，竹房あつ子，中田秀基，小口正人：姿勢推定ライブラリ OpenPose を用いた機械学習による動作識別手法の比較，第 81 回全国大会講演論文集，Vol. 2019, No. 1, pp. 275–276（オンライン），入手先 ⟨https://ci.nii.ac.jp/naid/170000179926/⟩ (2019).